AD-A095 397 RAND CORP SANTA MONICA CA F/G 9/2
DATA QUALITY INDICATORS AND THEIR USE IN DATA BASE SYSTEMS,(U)
MAY 80 R L PATRICK
UNCLASSIFIED RAND/P-6491 NL

END
DATE FILMED
3-81
DTIC

LEVEL II

(1)

(12) 13

DATA QUALITY INDICATORS AND THEIR USE IN DATA BASE SYSTEMS

(11) Robert L. Patrick

(11) May 1980

(14) RAND/P-6491

P-6491

The Rand Paper Se[illegible]

Papers are issued by The Rand Corporation as service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

DATA QUALITY INDICATORS AND THEIR USE IN DATA BASE SYSTEMS

Robert L. Patrick

DEFINITION

A Data Quality Indicator (DQI) is a descriptor used in computer file systems to record the quality attribute of the data. A DQI can be maintained at the file level to describe the quality of the file, at the record level to describe the quality of a record, or at the field level to describe the quality associated with a specific occurrence of a data element.

Data Quality Indicators are process time variables and their settings can determine which values participate in a computation and how that computation proceeds.

INTRODUCTION

One flaw in the computer based information systems of today lies with their inability to explicitly deal with uncertainty. As will be shown below, today's computer systems deal with error and uncertainty implicitly. In the past these implicit treatments have occasionally proven inadequate and other methods have been used. The Data Quality Indicator is sufficiently promising that it may become a standard feature in data based management systems of the future.

In its simplest form, the DQI is held in a one byte field immediately adjacent to the value whose quality it describes. When the value is read or written or moved, the accompanying quality indicator is also moved so the pair of fields are never separated. When the value is processed, the processing program can optionally process the quality indicator or ignore it. Thus in a gross sense the value field becomes one byte longer.

Quality information is coded and recorded in the indicator. In a simple system four quality states would be defined:

0 - value missing

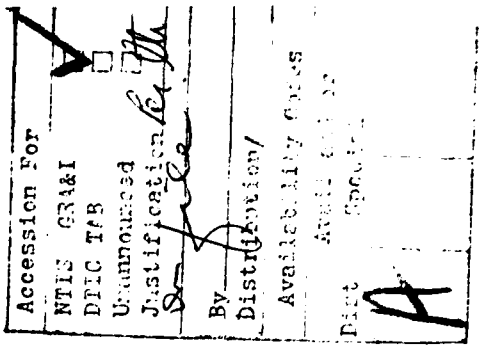
1 - value present but estimated

2 - value present but suspect

3 - value present and reliable.

For some statistical purposes all values with any non-zero quality indicator could be used. In other instances only values with a quality indicator of 3 would be processed. For control purposes the population of values with indicator codes of 1 and 2 would describe the work remaining to clean up the file.

This simple example describes the concept adequately. The remainder of this paper describes how quality indicators have been used to date and how this concept can lead to the solution of a major data processing problem that lies in the foreseeable future.



BACKGROUND

Traditional batch data processing causes a file of data to be created and held on magnetic tape. As changes are applied to that file, magnetic tape technology causes a new tape to be written. All large batch processing applications running today are based on that process. The controls over the process are supplied by a production control section. When a job is to be run, production control personnel determine which version of the magnetic tape is to be used for that specific run. If a file is being established, the latest copy of the tape is always the one to be used and that tape accumulates the union of all changes made. In this case the file naming convention supplies a single name to the basic file and version numbers are created by the update process.

Note that this process creates several copies of the same data which are nearly duplicates of one another. Furthermore, the production control methodology requires a short push-down list to be maintained so the most recent copy is constantly identified.

The same mechanism is used by the financial community for an entirely different purpose. Books of accounts are cyclic in nature. There are weekly payrolls, monthly bank statements, and annual taxes. At any one time the cost accountants deal with three versions of their basic monthly files:

(a) The open file accumulates the charges for an accounting period and generally is not used for any other purpose while in this state. Thus, it cycles on itself like the simple example described in the previous paragraphs.

(b) Once the appropriate calendar period has passed, an event occurs known as "closing the books." This event is really a series of steps where the first step involves opening a new file for the accumulation of the next period's expense data. Simultaneously the file for the period just passed is closed to new input and has its use restricted to a small category of bookkeepers until it can be brought into balance. During this process, totals are calculated and reconciled against other totals separately developed from independent sources (the sum of the payrolls is compared against the bank account). Each error found is investigated, corrections are developed, the file is updated, and the new balance computed. When the file is finally in balance it is used for management reporting.

Note that this file differs from the open version since it is closed to new input, and it purports to cover all the incurred expense for a specific time period. The values are usually of the same format and the same length in both the open and the closed versions, although the closed version contains many fewer zero values than the open copy.

(c) After the monthly accounting files are closed, reconciled, and used for management reporting, they are archived. In-depth analysis is performed from these archived tapes to produce historical cost trends, profitability analyses, and Government reports.

At year end 14 tapes, containing a high degree of redundancy, would contain monthly accounting data. These would consist of 11 tapes for the past months, 2 tapes for the month being closed, and 1 tape collecting next month's data. These files would be structually identical and identically named with date controls to allow production control personnel to differentiate between the sequential versions. Note: these versions differ in two fundamental characteristics: data timeliness and quality.

In the last 10 years the inexpensive demountable disk files have caused some minor changes in the physical handling of data files without any major changes occurring to the logical control of the process. As disk storage became less expensive, master files were migrated from tape to disk. The first benefit was fewer tape mounts and less computer operator labor.

As the computer processing programs were changed to make use of direct access storage devices, the computer resources involved in file update were reduced since only the data requiring change was processed (as it was unnecessary to recopy the correct portions of the file each time a correction transaction was processed).

Later the reconcilation and balancing efforts were further streamlined through the use of on-line processing so transactions need not be batched for a massive overnight file update. On-line update allowed the reconcilation process to proceed faster although separate versions of the file corresponding to the separate accounting periods were still maintained.

This traditional processing flow has been altered by the introduction of Data Quality Indicators. The following section describes how DQIs have been used in two production systems.

CASE STUDIES

Data Quality Indicators have been in use for about 10 years. Consider the following problems and their solutions:

Problem No. 1: A distributed computer system was to be used for keyboarding, editing, and correction of some very complex input records. After the data was keyboarded it was queued until a supervisor ran a program to verify that the batch totals were correct. After the batch totals were correct a first stage edit program verified that the contents of each data field met the edit criteria for that field. After the fields were correct a second stage edit program verified that the combination of legitimate fields was in fact allowable.

After these three checks the data was transmitted to the host machine where the contents of the records and their related transaction codes were edited in the presence of the masterfile, e.g., a masterfile record with a matching key was required if the transaction called for an update. If all of these steps were successfully carried out, the record was accepted by the system and was reflected in the masterfile during the next update cycle.

Solution No. 1: One way to have designed this multi-stage edit process would have been to maintain separate queues at each stage of the process and to move records from queue to queue as appropriate after each stage of the process was completed.

Instead of separate queues, a Data Quality Indicator was appended to each record and one common queue was kept on the minicomputer and a second queue was kept on the host computer. When a record was originally keyboarded the quality indicator was set to zero. As it successfully passed each stage of edit the quality indicator was increased by one. When the data was transmitted from the minicomputer to the host computer the quality indicator on the records in the minicomputer was increased by one to indicate that the space was available for reallocation. This simplified programming, storage allocation, and control over the process.

Problem No. 2: A large social science research project was to gather data by interviewing a cross-section of individuals in their homes. Some of the important factual data was likely to vary in quality depending on the source of documents available during the interview, e.g., income read from a W-2 form is intrinsically more accurate than data remembered or estimated. Further, after the questionnaires were administered in the field they were manually edited, keyboarded, machine edited, and entered into a data base. Further, the original data from the field were to be retained in the file regardless of quality since change to the original input (except to eliminate transcription errors) inhibits methodological studies which evaluate the data-gathering process.

Solution No. 2: Two Data Quality Indicators were appended to each critical data value. Codes describing the source documents from which the data was extracted in the field were translated to set the first data quality indicator. Thus the day laborer who had been paid in cash and recollected his total income from a variety of sources appeared much lower on the data quality scale than a widow whose sole income was from monthly pension checks and who produced a year's worth of check stubs.

During the data cleaning process, the second data quality indicator was set. Data passing all tests received the highest quality code. Data failing one or more tests received a quality indicator which identified the process catching the error and the specific test failed.

Some extra codes were defined which allowed the data quality indicators to be changed as a result of in-depth research analysis. Sometimes the edits performed during the cleaning process were not adequate and sophisticated errors slipped through and got into the data base. Since several research activities were drawing from the same data base, arbitrary changes to the data could not be tolerated as they disturbed the run-to-run continuity the researchers desired to maintain. However, if sophisticated data errors were uncovered (the total income from all sources was less than the sum of expenses and services provided in kind), suspicious data flags were set in the records so subsequent researchers would not unknowingly include these data records in their sample for future research.

During the research process, investigators called for a subset of the data based on the data quality indicators so they could investigate the phenomena initially with data having a low uncertainty. Then after the research methodology had been proved out and they were familiar with the best behaved members of the sample, they relaxed their selection criteria and investigated the other members of the population who, while not so well-behaved, were more interesting.

CONJECTURES

Even in their primitive form, data quality indicators would be useful for law enforcement or credit data files. Thus a rumor off the street, a report from a usually reliable source, a record from a cooperating law enforcement agency, an observation by a local officer or facts determined in a court of law, could each be represented by separate data quality codes. Similarly in credit files a report from a cooperating credit agency, an individual's signed application, data from a credit grantor, or data from the person's bank, could be treated appropriately.

Simple statistics on the indicator codes provide a quality profile on the data file. If it were necessary to process data having several quality indicators imbedded within it, logical arithmetic on the indicators could be performed and summarized on the printed reports to provide the decision makers with the information needed to determine whether to take action, and what action to take.

As dynamic data dictionaries (those held on-line and used to control the processing of data) become commonplace, a quality histogram (count of data records at each quality level) could be included in the master description record for each data set so each person or process intending to use the file could be apprised of the status of the file before processing was initiated.

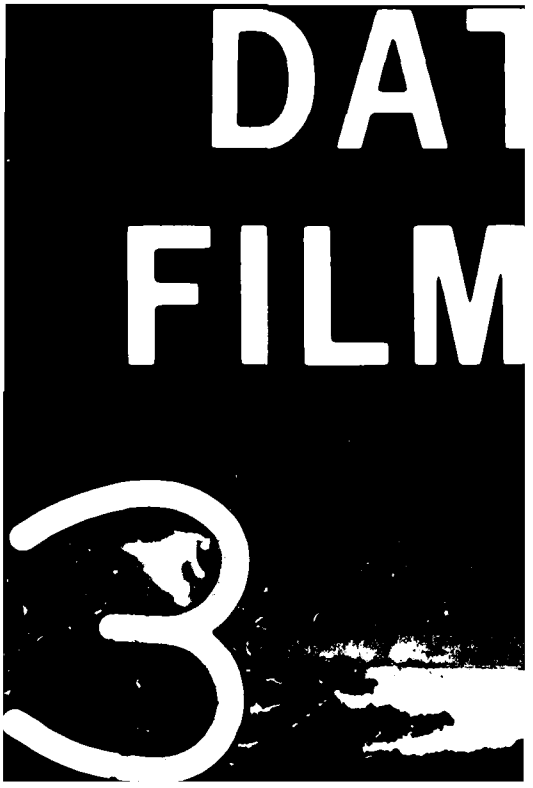
The future may see a date field and a quality indicator as two commonplace secondary attributes in every very large file system. The date would record when the item (file, record or field) was last updated and the quality indicator would record the known quality on that date. Simple processing involving these two attributes would correct two of the frequently occurring errors in the handling of files containing personal data, i.e., using obsolete data or using data whose quality was not appropriate for its intended purpose.

FUTURE PROSPECTS

Computer technology is leaning towards large, inexpensive, on-line bulk files. Recent IBM announcements offer 571 megabytes of disk for less than $1000 a month. When measured by capacity per dollar, that is a 428% improvement in only 10 years. Some reliable technology watchers speculate that this trend will continue for several more cycles. Thus, it can be concluded that all of the small- and medium-sized files kept by a business enterprise and some of the large files will be maintained permanently on-line in the foreseeable future.

However, to keep several versions of each file on-line containing almost duplicate data will be an extravagance that few will choose to afford. Therefore, some method needs to be developed which will allow the technology to be exploited without unreasonable costs in storage or performance being paid.

Further, the time is appropriate for an in-depth review of the production control function so it can be changed to accommodate terminal users who wish to safely interact with on-line information files. The Data Quality Indicator, together with appropriate algorithms for dynamically processing these indicators and their related data, appears to offer some hope in solving a critical problem related to the evolving technology.


DAT
FILM